DOCUMENT RESUME

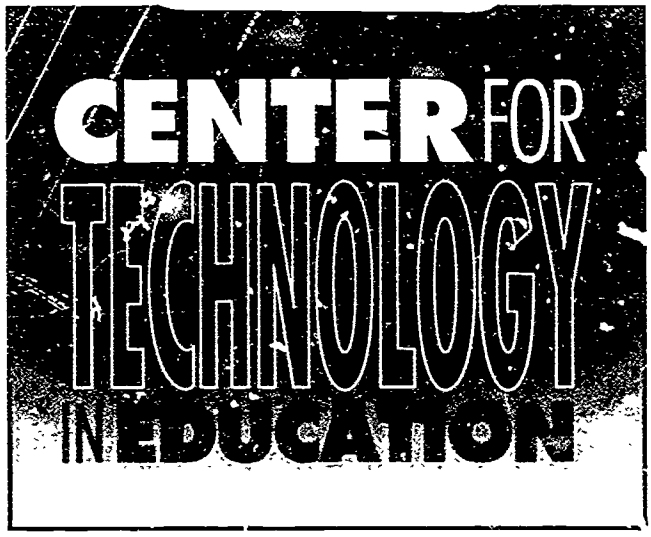
ED 325 484 TM 015 668

AUTHOR Frederiksen, John R.; Collins, Allan
TITLE A Systems Approach to Educational Testing. Technical Report No. 2.
INSTITUTION Center for Technology in Education, New York, NY.
SPONS AGENCY Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE Jan 90
CONTRACT OERI-1-135562167-A1
NOTE 12p.
PUB TYPE Viewpoints (120) -- Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS Cognitive Development; Cognitive Tests; *Educational Assessment; Educational Change; Elementary Secondary Education; Outcomes of Education; Skill Development; *Student Evaluation; *Systems Approach; Test Construction; Testing Problems; *Test Validity

ABSTRACT

The validity of educational tests used as critical measures of educational outcomes within a dynamic system is discussed. Validity becomes a problem if an educational system adapts itself to the characteristics of the outcome measures. The concept of systematically valid tests is introduced; these tests induce curricular and instructional changes in education systems and learning strategy changes in students that foster the development of the cognitive traits the tests are designed to measure. Two characteristics are analyzed that contribute to or detract from a testing system's systemic validity: (1) use of direct rather than indirect cognitive assessment; and (2) the degree of subjectivity or judgment required to assign a score to represent the cognitive skill. These characteristics are then applied in developing design principles for creating systematically valid testing systems. These principles are illustrated in the design of a student assessment system that includes the means of teaching the process of assessment to system users. A list of 29 references is attached. (SLD)

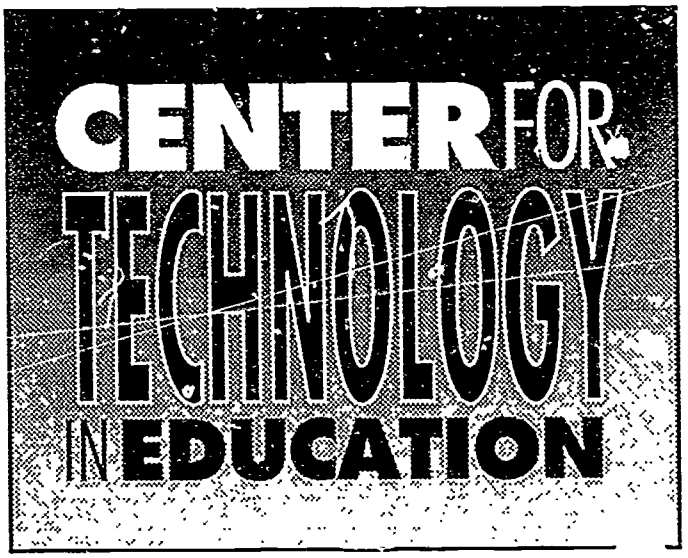


A Systems Approach to Educational Testing

John R. Frederiksen & Allan Collins
Bolt Beranek and Newman

BANK STREET COLLEGE OF EDUCATION
610 WEST 112th STREET
NEW YORK, NY 10025

Technical
Report
Series



Technical Report No. 2

A Systems Approach to Educational Testing

John R. Frederiksen & Allan Collins
Bolt Beranek and Newman

January 1990

Tech. Rep. No. 2

A SYSTEMS APPROACH TO EDUCATIONAL TESTING

John R. Frederiksen and Allan Collins
Bolt Beranek and Newman

Abstract

Our concern in this paper is with the validity of educational tests when they are employed as critical measures of educational outcomes within a dynamic system. The problem of validity arises if an educational system adapts itself to the characteristics of the outcome measures. We introduce the concept of systemically valid tests as ones that induce curricular and instructional changes in education systems (and learning strategy changes in students) that foster the development of the cognitive traits that the tests are designed to measure. We analyze some general characteristics that contribute to or detract from a testing system's systemic validity, such as the use of direct rather than indirect assessment. We then apply these characteristics in developing a set of design principles for creating testing systems that are systemically valid. Finally, we provide an illustration of the proposed principles by applying them to the design of a student assessment system. This design example addresses not only specifications for the tests, but also the means of teaching the process of assessment to users of the system.

There are enormous stakes placed on students' performance on educational tests. And there are, consequently, enormous pressures on school districts, school administrators, teachers, and students to improve scores on tests. These pressures drive the educational system to modify its behavior in ways that will increase test scores (Darling-Hammond & Wise, 1985; Madaus, 1988). The test scores, rather than playing the role of passive indicator variables for the state of the system, become the currency of feedback within an adapting educational system. The system adjusts its curricular and instructional practices, and students adjust their learning strategies and goals, to maximize the scores on the tests used to evaluate educational outcomes, and this is particularly true when the stakes are high (Corbett & Wilson, 1988). Thus, for example, if a reading test emphasizes certain skills, such as knowledge of phonics, then these become the skills that will receive emphasis in the reading curriculum.

Our concern in this paper is with the validity of educational tests within such a dynamic system. To introduce tests into a system that adapts itself to the characteristics of tests poses a particular challenge to their validity and calls into question many of the current practices in educational testing. That challenge to validity has to do with the effects of the instructional changes engendered by the use of the test and whether or not they contribute to the development of the knowledge and/or skills that the test purportedly measures. This extension of the notion of construct validity of a test to take into account the effects of instructional changes brought about by the introduction of the test into an educational system we shall refer to as the systemic validity of a test. A

Published in Educational Researcher, Vol. 18, No. 9, pp. 27-32 (1989, December).

systemically valid test is one that induces in the education system curricular and instructional changes that foster the development of the cognitive skills that the test is designed to measure. Evidence for systemic validity would be an improvement in those skills after the test has been in place within the educational system for a period of time.

Given this challenge to test validity due to systemic effects, the question we must take up has to do with whether there are any general characteristics of a system of testing that can be identified as either contributing to or detracting from a test's systemic validity. In our analysis, we shall identify a number of characteristics that contribute to systemic validity. We shall then apply these principles in developing a set of design principles for an alternative form of testing system that is systemically valid—one that we believe will drive the educational system toward practices that will lead to improvements in the underlying knowledge and skills that tests are seeking to measure. Finally, we shall provide an illustration of the proposed principles, in the context of a student assessment system. (Elsewhere, we have applied the design principles to teacher assessment, Collins & J. R. Frederiksen, 1989).

Educational Systems as Dynamic Systems

The measures that educators choose to use in assessing outcomes provide one important form of feedback that determines how the system will modify its future operation. Schoenfeld's (in press) observations of the teaching of one of the most successful math teachers in New York State precisely illustrates our point. Students of geometry in the state of New York must all pass a statewide Regents' Exam that has become, in no uncertain terms, the goal of instruction. Scores on the test are used to judge students, teachers, and school districts. In geometry, the exam includes as a major component a required proof (chosen from a list of a dozen theorems) and also a construction problem (in which tools such as a straightedge and a compass are used to "construct" a figure with specified properties). In the scoring of the proofs, students are expected to reproduce all the steps of the proof in a two-column form, listing each proof step and a justification for that step. In the construction problem, they are not required to give justifications for the steps of the construction, but are graded on whether the construction has all of the required arcs and lines and how accurately they are drawn. Schoenfeld found that these characteristics of the Regents Exam have completely subverted the way the teacher taught geometry. Instead of teaching students how to generate proofs, the teacher had students memorize the steps for each of the 12 proofs that might be on the exam. In their constructions, the students were taught how to carry them out neatly. The students were thus able to pass the geometry part of the Regents' Exam with flying colors, but they did not learn how to reason mathematically.

This example illustrates how the systemic validity of a test is dependent on the specification of the construct the test is taken to measure, which is in turn related to the goals of teaching and learning. If the goal of teaching geometry is to be able to reproduce formal proofs and to develop flawless constructions, then the Regents' geometry test can be said to be systemically valid. However, if the goal is to assess how students can develop proofs and use constructions as tools for mathematical exploration, then the test cannot be said to be systemically valid, because its use has engendered instructional adaptations that do not contribute to the development of these cognitive skills. A test's validity cannot be evaluated apart from the intended use of the test (Messick, 1988).

In the absence of feedback and adaptation to the test, the Regents' test and tests like it may provide an adequate indication of students' knowledge, because most representative geometry items will correlate highly with one another and the use of one or another particular set of test items will not result, therefore, in any gross misclassification of test takers. However, the requirement of systemic validity creates a much more stringent standard for the construction of tests, for it requires us to consider evolutions in the form and content of instruction and students' learning engendered by use of the test. That is, will instruction that focuses on the skills and problem formats represented in tests promote the ability of students to engage, in the present case, in authentic mathematical investigations and problem solving? There are several reasons why we believe that it will not.

1. If a test emphasizes isolated skill components and items of knowledge, instruction that seeks to increase test scores is likely to emphasize those skill components rather than higher level processes (N. Frederikse., 1984; Resnick & Resnick, in press).

2. Instruction that seeks to develop specialized test-taking strategies (e.g., in taking a multiple choice test, trying to eliminate one or more of the response

alternatives and then guessing) will not improve domain knowledge and skills.

3. Time and effort spent in directly improving test scores in these ways will displace other learning activities that could more directly address the skills and learning goals the test was supposed to be measuring in the first place.

4. Students will direct their study strategies toward those skills (such as memorization) that are represented on the tests—and that appear to be valued by educational institutions—rather than toward the use of cognitive skills and knowledge in solving extended problems.

One solution to the problem of low systemic validity would be, of course, to disallow the development of any instruction aimed explicitly at improving scores on the test. Such an approach, however, would deny to the educational system the ability to capitalize on one of its greatest strengths: to invent, modify, assimilate, and in other ways improve instruction as a result of experience. No school should be enjoined from modifying its practices in response to their perceived success or failure. Nor should students be prevented from optimizing their study so as to carry out the kinds of problem solving valued within their course of study. Yet if these strategic modifications in teaching and learning are to be based on test scores, then their efficacy will depend crucially on the systemic validity of the tests that are used. We are left, therefore, with the alternative solution to the problem: to encourage the inventiveness and adaptability of educational systems by developing tests that *directly reflect and support the development of the aptitudes and traits they are supposed to measure.*

Characteristics of Systemically Valid Tests

There are two dimensions or characteristics of tests that have a bearing on their usefulness as facilitators of educational improvement. These are (a) the directness of cognitive assessment, and (b) the degree of subjectivity or judgment required in assigning a score to represent the cognitive skill.

In *indirect tests*, an abstract cognitive skill is measured by evaluating less abstract, more directly observable features of performance that are known (or theoretically expected) to be highly correlated with the abstract skill. For example, verbal aptitude, a construct that might be defined as "the ability to formulate and express arguments in verbal form," is measured using tests of vocabulary knowledge or verbal analogies. In *direct tests*, the cognitive skill that is of interest is directly evaluated as it is expressed in the performance of some extended task. An example would be to rate the coherence of an argument in a legal brief.

The *degree of subjectivity* of a test refers to the degree to which judgment is used in assigning a score to a student's test performance. *Objective tests* use simple, algorithmic scoring methods such as counting the number of items correct. *Subjective tests*, on the other hand, require judgment, analysis, and reflection on the part of the scorer in the assignment of a score. Because the scoring algorithms of objective tests are simple, the item formats of such tests are usually constructed to invoke unitary responses, such as selecting one from a set of multiple-choice response alternatives or writing a single word, phrase, or number. Subjective tests do not necessitate this restriction on the form of response and typically allow more extended responses to a test item, such as the writing of an essay. Drew Gitomer (personal communication, May 8,1989) has pointed out that in objective tests, there is a low degree of inference required at the item-scoring level, but a much higher degree of inference required when items are aggregated using a psychometric model (e.g., item response theory, factor analysis) to produce a scale representing a particular construct. Subjective tests require, in contrast, more judgment and expertise in scoring at the item level, but very little inference at the level of summarizing item level scores. In educational testing, objective tests are generally preferred because they reduce the scoring task to a simple, objective scoring algorithm such as a tallying of correct answers. Benefits of such objective tests are the reliability of scoring, the lack of potential biases that might affect score assignments, and the ease and economy of algorithmic scoring.

Problems with using objective tests. We believe that one pays a very high price in reduced systemic validity for using objective tests. This is due to the fact that the desire for objective tests leads to tests that are indirect, and indirect tests often have problems of systemic validity. For example, in teacher assessment, competency can be assessed using tests of teachers' knowledge (domain knowledge and pedagogical knowledge) and basic skills (e.g., reading and mathematics). However, while such knowledge may be associated with or even necessary for effective practice as a teacher, it does not provide direct evidence of such practice, nor will developing such knowledge

ensure more effective teaching. Similar remarks can be made about tests of factual knowledge as a measure of accomplishment at the end of a course in history or tests of vocabulary knowledge as a measure of the capacity to do college work. In general, objective tests emphasize low-level skills, factual knowledge, memorization of procedures, and isolated skills, and these are aspects of performance that correlate with but do not constitute the flexible, high-level skills needed for generating arguments and constructing solutions to problems (N. Frederiksen, 1989; Resnick & Resnick, in press). Use of objective tests thus leads to teaching strategies that emphasize the conveying of information and to student learning strategies that emphasize memorization of facts and procedures, rather than learning to generate solutions to problems—including novel problems that occur in "real life" contexts. N. Frederiksen (1984) has termed this effect of tests on the content of instruction "the real test bias."

In some cases, it may be possible to construct objective tests that are direct measures of important cognitive constructs, such as identifying mental models in physics (Clement, 1982; McCloskey, Caramazza, & Green, 1980; McDermott, 1984; White, 1983) or assessing creativity in scientific problem solving (N. Frederiksen, 1978). It may also be possible to use techniques of artificial intelligence to build relatively detailed models of students' knowledge on the basis of extended examples of their problem solving (Anderson, Boyle, & Reiser, 1985; Clancey, 1983; J. R. Frederiksen & White, 1989; Johnson & Soloway, 1985; Sleeman & Brown, 1982). Although it is worthwhile to continue efforts to develop objective tests of important cognitive outcomes of learning, in general the state of the art does not permit objective tests for directly measuring higher order thinking skills, problem-solving strategies, and metacognitive abilities involved in tasks such as teaching, writing, constructing a historical argument, and "doing" mathematics. Thus we believe that it is important to consider some of the advantages of subjective, direct assessment of such high-order cognitive skills.

Advantages of direct tests. Direct tests attempt to evaluate a cognitive skill as it is expressed in the performance of extended tasks. Such measures are systemically valid, because instruction that improves the test score will also have improved performance on the extended task and the expression of the cognitive skill within the task context. In figure skating and gymnastics, for example, measures of traits such as technical merit and artistic impression are assigned by judges based on an extended program that is developed and performed by the athlete.

In educational testing, a particularly good example of this approach (and one that has been seminal in influencing our thinking) is the primary trait system for scoring writing tasks that was developed by the National Assessment of Educational Progress (NAEP) (Mullis, 1980). The purpose of the NAEP assessment was to measure whether a piece of writing is successful or unsuccessful in achieving a particular purpose. The student is given a writing assignment with a particular goal, such as writing a letter to the chairman of the school board on the advisability of instituting a 12-month school year. To evaluate such writing, a set of primary traits was developed that are important for successfully achieving the goal of the writing assignment. For example, one primary trait, persuasiveness, involves the presentation of a set of logical and compelling arguments. The completed writing exercise is rated on a set of such primary traits, using a simple 4-point scale for each. For example, persuasiveness is rated as follows: "1" for a paper containing no reasonable argument, "2" for a paper having one or two poorly thought out arguments, "3" for a paper containing several logically thought out reasons, and "4" for a paper containing in addition a number of compelling details (Mullis).

Basing educational assessment on such subjective scoring requires that scorers understand the scoring categories and be taught how to use them reliably. This in turn necessitates building a library of exemplars of student work representing different levels of the desired primary traits. This library is then used to train scorers to assess the traits. In the case of the NAEP writing assessment, for each writing exercise, exemplars of texts scored in each category are provided. In addition, a detailed rationale is included for each exemplar explaining why the particular score has been assigned. Assessors study these exemplars and practice scoring until they have internalized the criteria and can rate primary trait performance reliably in a variety of task contexts. In the NAEP primary trait assessment of writing, a typical interscorer agreement of 91%-95% was achieved. Moreover, studies have shown that individual, remote scorers, following calibration (Braun, 1986), can provide scores that approach quite closely the values derived using standardized scoring methods (Breland & Jones, 1988).

It would be difficult to justify the cost of develop

ing these training materials if they were to be used only to train professional assessors. However, there is another use to which they can be put: *The training materials can become the medium for communicating to teachers and students the critical traits to look for in good writing, good historical analysis, and good problem solving.* The library of exemplars can be viewed as a set of "case studies" that can be used by teachers to make their students aware of the nature of expert performance, or as Wolf puts it, to help them "develop a keen sense of standards and critical judgment" (1987, p. 26). Using them, students can learn to assess their own work in the same way that their teachers will judge it. They can, for example, learn to recognize critical traits in their writing and to carry this awareness along with them as they carry out their assignments. The assessment system provides a basis for developing a metacognitive awareness of what are important characteristics of good problem solving, good writing, good experimentation, good historical analysis, and so on. Moreover, such an assessment can address not only the product one is trying to achieve, but also the process of achieving it, that is, the habits of mind that contribute to successful writing, painting, and problem solving (Wiggins, 1989). We believe that building such awareness will lead to genuine improvements in the cognitive traits on which the assessment system is based.[1] We argue, therefore, that adopting subjective, direct assessment is a good way to increase the systemic validity of a testing system.

Principles for the Design of Systemically Valid Testing

Our plan for the design of a systemically valid testing system has three major aspects: (a) the components of the testing system; (b) the standards to be sought in the design of the system; and (c) the methods by which the system encourages learning. A general outline of the design specification will be presented in this section. In the subsequent section, we will illustrate the applications of this design for a student assessment system.

Components of the Testing System

The testing system we envision has four major components: a set of tasks, a specification of primary traits to be assessed, a library of exemplars of performances on each task, and a training system for teaching how to score the primary traits.

Set of tasks. The tests should consist of a representative set of tasks that cover the spectrum of knowledge, skills, and strategies needed for the activity or domain being tested. For example, in student assessment, if there is a set of basic problem-solving skills we think students should acquire, these skills must be called for in the tasks given. The tasks might be constructed as in the assessment of figure skating, a set of compulsory tasks plus a set of elective tasks, so that testees can demonstrate both their basic abilities in compulsory tasks and their planning and creativity in elective tasks. The tasks should be authentic, ecologically valid tasks in that they are representative of the ways in which knowledge and skills are used in "real world" contexts (Brown, Collins, & Duguid, 1989, Wiggins, 1989).

Primary traits for each task and subprocess. The knowledge and skills used in performing any task may consist of distinct subprocesses. For example, teaching might be broken down into planning, classroom practice, and evaluating students' work, each of which requires somewhat different talents. These subprocesses need to be assessed independently so that test takers will direct their efforts to doing well in all phases of the task domain being tested. Each subprocess must be characterized by a small number of *primary traits* or characteristics that cover the knowledge and skills necessary to do well in that aspect of the activity. The traits should cover both process and products and should include planning and reflection. For example, in writing, processes might include note taking, outlining, drafting, and revising. The primary traits for expository writing might be clarity, persuasiveness, memorability, and enticingness (Collins & Gentner, 1980). (The specific traits may differ for different processes and products.) The primary traits chosen should be ones that the test takers should strive to achieve, and thus should be traits that are learnable. The small number is necessary to focus the test taker's learning. The particular traits chosen for any task domain are not too critical, as long as they cover the skills that are judged to be important and they are learnable. In other words, we believe that the testing approach is robust over different sets of primary traits.

A library of exemplars. In order to ensure reliability of scoring and learnability, it is important that for each task there be a library of exemplars of all levels of performance for each primary trait assessed in the

test. The library should include exemplars representing the different ways to do well (or poorly) with respect to each trait. It should also include critiques of each sample performance, so that it is clear how the performance was judged. The library should be accessible to all, and particularly to the testees, so that they can learn to assess their own performance reliably and thus develop clear goals to strive for in their learning.

A training system for scoring tests. There are three groups that must learn to score test performance reliably: (a) the administrators of the testing system, who develop and maintain the assessment standards (i.e., master assessors); (b) the coaches in the testing system whose role is to help test takers to perform better; and (c) the test takers themselves, who must internalize the criteria by which their work is being judged. The master assessors are charged with defining the criteria, ensuring that test performance can be scored reliably, and training coaches to score performances. The coaches work with the test takers to teach them self-assessment.

Standards

Standards must be developed for the testing system that include the following:

Directness. From a systems point of view, we have seen that it is essential that whatever knowledge and skills we want test takers to develop be measured directly. Sometimes this may require measuring a process, sometimes a product, and sometimes both. In either case, any indirectness in the measure will lead to a misdirection of learning effort by test takers to the degree that it matters to them to do well on the test.

Scope. The test should cover, as far as possible, all the knowledge, skills, and strategies required to do well in the activity. To the degree that any knowledge or skills are left out, test takers will direct their learning efforts to only part of what is required of them.

Reliability. We think that the most effective way to obtain reliable scoring that fosters learning is to use primary trait scoring borrowed from the evaluation of writing. Developing a primary trait system for any test involves the same steps that were used by NAEP in applying it to writing.

Transparency. The terms in which the test takers are judged must be clear to them if a test is to be successful in motivating and directing learning (Wiggins, 1989). In fact, we argue that the test must be transparent enough so that they can assess themselves and others with almost the same reliability as the actual test evaluators achieve.

Methods for Fostering Improvement on the Test

The testing system should not only employ forms of assessment that enhance learning, but it should also include specific methods designed to foster such learning. These include the following.

Practice in self-assessment. The test takers should have ample opportunity to practice taking the test and should have coaching to help them assess how well they have done and why. This kind of reflection on performance (Collins & Brown, 1988) is made possible by recording technologies such as videotape and computers. The assistance of a coach, who has internalized the testing standards, is critical to helping the test takers see their performance through others' eyes.

Repeated testing. Although it may be necessary to have the test administered at only a few times during a year, it is still important to encourage students to take the test multiple times to encourage striving for improvement. If what is measured by the test is important to learn, then the test should not be taken once and forgotten. It should serve as a beacon to guide future learning.

Feedback on test performance. Whenever a person takes the test, there should be a "rehash" with a master assessor or teacher. This rehash should emphasize what the testee did well and poorly on, and how performance might be improved. It should preferably involve a master assessor so that the institutionalized standards will be clear to the test taker.

Multiple levels of success. There should be various landmarks of success in performance on the test, so that students can strive for higher levels of performance in repeated testing. The landmarks or levels might include such labels as "beginner," "intermediate," and "expert" to motivate attempts to do better.

Student Assessment

The system we envision involves developing a number of extended tasks or projects that students would carry out to demonstrate their mastery of courses they are taking, such as history or physics. We can illustrate the approach with two structured tasks that might be given to students in American history and physics. For history, a task might be as follows: "At the beginning of World War II, the United States was divided as to whether to enter the war or to stay neutral. Pick three presidents in history, other than Franklin Roose-

velt, who you think would have taken different positions on the issue, and write a 2-minute speech of each to the American public on what should be done in that situation." These speeches might then be delivered and recorded on videotape, with questions following from other students as in a press conference. For physics, the task might be to design a set of activities using a Dynaturtle (diSessa, 1982; White, 1984) that would help younger students learn to understand Newton's Laws of Motion. (A Dynaturtle is an object in a computer simulation that operates in a frictionless, gravity-free environment, and is controlled like a spaceship.) These are examples of the kind of extended tasks that students could be given to demonstrate their understanding of history or science. A variety of such tasks could be provided to teachers for use in assessment, or teachers could construct their own tasks following a set of task specifications that are provided to them. In general, the tasks to be included within an assessment system would vary from structured tasks that measure students' understanding of critical concepts or skills to open-ended tasks that allow students to demonstrate special knowledge and creativity. Ideally, these tasks would be fully integrated within a course, rather than serving as accessories to the course.

Scoring Student Performance

Students would be evaluated on the tasks in terms of a set of primary traits. Examples of primary traits that could be used are (a) clarity of expression, (b) creativity, (c) depth of understanding or thoroughness, (d) consideration of multiple perspectives, and (e) focus or coherence. The particular traits chosen are, again, not critical so long as they cover the desired qualities and direct students' efforts appropriately. The primary traits would cover both process and products, and also might be applied to different phases of an assessment task, such as planning, presentation, and revision.

To implement the assessment system, it is important to build a library of exemplars of students working on a variety of tasks, covering all the major subject areas. This library would be embodied in paper, videotapes, and computer traces. For example, paper records might include notes, outlines, and multiple drafts of articles written. Videotapes might record students discussing their initial plans, making presentations, answering questions, or performing dramatic scenes. Computers might record document preparation and revision or students' solutions to problems such as the physics activity described above. Each of these exemplars should also contain a critique of the performance by master assessors in terms of the set of primary traits chosen for evaluating students.

The administration for such a system could be centered at the school, district, state, or even national level. There would have to be a group of master assessors who are responsible for developing the set of traits, the criteria for scoring, and the library of exemplars. They would also be responsible for showing teachers how to evaluate student performance, and in fact testing teachers to make sure that they have internalized the evaluation criteria. Teachers would function as coaches to the students as they practiced different tasks, to help them internalize the criteria by which they are judged. Ideally, students would learn how to critique their own and each other's performances in terms of the primary traits adopted.

Addressing Different Audiences

A major problem in student assessment is that the test scores generated have to address the needs and desires of many different audiences. Colleges need to know whether the student meets their admission standards. Teachers want to know what students have learned and failed to learn. Parents and students want to know how the student is doing relative to some standard. Administrators want to know how well different teachers and schools are succeeding. All of these different needs have to be balanced in setting up an assessment system.

Because colleges are a major constituency for student assessment, the criteria for evaluating students in each subject should be developed in conjunction with college admissions officers, who have ideas about what are essential knowledge and skills for admission. (For students in vocational courses, criteria should be developed in consultation with businesses and other potential employers and with licensing boards.) These same criteria should suffice for parents, students, and teachers, since they are the outcome measures that are valued by colleges or future employers, and are therefore ecologically valid measures of performance that are judged to be important in "real world" tasks.

A Changing Role for Testing Organizations

Lest the proposal for a systemically valid testing system we have made seem overly visionary, we shall

examine briefly the practical side of implementing such a system. We believe that the efficiency in current testing practices is greatly outweighed by the cost of using a system that has low systemic validity—one that has a negative impact on learning and teaching. The goal of assessment has to be, above all, to support the improvement of learning and teaching. To accomplish this, major changes must occur in the role and function of testing organizations. In the future, they will retain their important role as developers of assessment tools, and they will, as now, be responsible for setting scoring standards and practices. However, they will have to assume some new responsibilities: (a) they must develop materials for use in teaching the assessment techniques, not only to master assessors within schools and school districts, but also to teachers and students; and (b) they must take responsibility for ensuring that the assessment standards are assimilated and maintained by these new groups of assessors. The big difference is that the practice of assessment will no longer be confined to the testing organizations; it will become more decentralized, as teachers and students are taught to internalize the standards of performance for which they are to strive.

We end with some caveats. Clearly, much research needs to be done to test the assumptions on which our proposal is based: Can primary traits be assessed reliably on a common scale when the particular tasks that test takers carry out may vary? Does an awareness of primary traits help students to improve performance on projects and teachers to become more effective in the classroom? Can a consensus be reached on what are appropriate primary traits for different domains and activities? Can scoring standards be met when assessment is decentralized? These and other questions should become the basis of a concerted research effort in support of a new, systemically valid system of educational testing.

Notes

This work was supported by the Center for Technology in Education under Grant No. 1-135562167-A1 from the Office of Educational Research and Improvement, U.S. Department of Education, to Bank Street College of Education. We would like to thank Norman Frederiksen, Drew Gitomer, Robert Glaser, and Ray Nickerson for their thoughtful comments on an earlier draft of the paper.

1. A critical assumption is that scorers can learn to recognize and reliably assess primary traits, not only in the particular tasks used in the library of exemplars, but in other tasks for which the trait is relevant. Although there is evidence bearing on these assumptions in the assessment of writing (Breland & Jones, 1988), further work will be required to check its validity for the specific primary traits that are to be the goal of assessment.

References

Anderson, J. A., Boyle, C. F., & Reiser, B. J. (1985). Intelligent tutoring systems. *Science, 228*, 456-68.

Braun, H. (1986). *Calibration of essay readers* (Report No. RR-86-9). Princeton, NJ: Educational Testing Service.

Breland, H. M., & Jones, R. J. (1988). *Remote scoring of essays* (Report No. 88-4). Princeton, NJ: Educational Testing Service.

Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher, 18*(1), 32-42.

Clancey, W. (1983). Guidon. *Journal of Computer-Based Instruction, 10*(1 & 2), 8-15.

Clement, J (1982). Students' preconceptions in elementary mechanics. *American Journal of Physics, 50*, 66-71.

Collins, A., & Brown, J. S. (1988). The computer as a tool for learning through reflection. In H. Mandl & A. Lesgold (Eds.), *Learning issues for intelligent tutoring systems* (pp. 1-18). New York: Springer.

Collins, A., & Gentner, D. G. (1980). A framework for a cognitive theory of writing. In L. W. Gregg & E. R. Steinberg (Eds.), *Cognitive processes in writing* (pp. 51-72). Hillsdale, NJ: Erlbaum.

Collins, A., & Frederiksen, J. R., (1989). *Five traits of good teaching: Learning, thinking, listening, involving, helping.* Unpublished report, BBN Laboratories, Cambridge, MA.

Corbett, H. D., & Wilson, B. (1988). Raising the stakes in statewide mandatory minimum competency testing. *Politics of Education Association Yearbook*, 27-39.

Darling-Hammond, L., & Wise, A. (1985). Beyond standardization. State standards and school improvement. *Elementary School Journal, 85,* 315-336.

diSessa, A. (1982). Unlearning Aristotelian physics. A study of knowledge-based learning. *Cognitive Science, 6,* 37-76.

Frederiksen, J. R., & White, B. Y. (1989). Intelligent tutors as intelligent testers. In N. Frederiksen, R. Glaser, A. Lesgold, & M. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 1-25). Hillsdale, NJ. Erlbaum.

Frederiksen, N. (1978). *Assessment of creativity in scientific problem solving* (Research Memorandum RM-78-9). Princeton, NJ: Educational Testing Service.

Frederiksen, N. (1984). The real test bias. *American Psychologist, 39*(3), 193-20...

Frederiksen, N. (1989). Introduction. In N. Frederiksen, R. Glaser, A. Lesgold, & M. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. vii—xv). Hillsdale, NJ: Erlbaum.

Johnson, W. L., & Soloway, E. (1985). PROUST: An automatic debugger for Pascal programs. *Byte, 10*(4), 179-190.

Madaus, G. (1988). The influence of testing on the curriculum. In L. Tanner (Ed.), *Critical issues in curriculum. 87th Yearbook of the NSSE, Part 1.* Chicago: University of Chicago Press.

McCloskey, M., Caramazza, A., & Green, B. (1980). Curvilinear motion in the absence of external force . Naive beliefs about the motion of objects. *Science, 210,* 1139-1141.

McDermott, L. C. (1984). Research on conceptual understanding in mechanics. *Physics Today, 37,*24-32.

Messick, S. (1988). Validity. In R. L. Linn (Ed., *Educational measurement* (3rd ed., pp. 13-103). New York. Macmillan.

Mullis, I. V. S. (1980). *Using the primary trait system for evaluating writing.* National Assessment of Educational Progress Report. Denver, CO. Education Commission of the States.

Resnick, L. B., & Resnick, D. P. (in press). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford & M. C. O'Connor (Eds.) *ure assessments. Changing views of aptitude, achievement, and instruction.* Boston: Kluwer.

Schoenfeld, A. H. (in press). On mathematics as sense-making: An informal attack on the unfortunate divorce of formal and informal mathematics. In D. N. Perkins, J. Segal, & J. Voss (Eds.), *Informal reasoning and education.* Hillsdale, NJ: Erlbaum.

Sleeman, D., & Brown, J. S. (Eds.). (1982). *Intelligent tutoring systems.* New York: Academic Press.

White, B. Y. (1983). Sources of difficulty in understanding Newtonian dynamics. *Cognitive Science, 7*(1), 41-65.

White, B. Y. (1984). Designing computer activities to help physics students understand Newton's laws of motion. *Cognition and Instruction, 1,* 69-108.

Wiggins, G. (1989, May). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan,* 703-713.

Wolfe, D. P. (1987, December). Opening up assessment. *Educational Leadership,* 24-29.

Appendix 16

END

U.S. Dept. of Education

Office of Education
Research and
Improvement (OERI)

ERIC

Date Filmed

March 29, 1991